DOCUMENT RESUME

ED 208 035 TM 810 732

AUTHOR Thompson, Pat
TITLE The Consistency of Title I Evaluation Results for Commonly Used Test Batteries.
PUB DATE Apr 81
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Comparative Analysis; *Diagnostic Tests; Educational Assessment; Elementary Education; *Mathematics; *Reading Achievement; *Standardized Tests; Test Use
IDENTIFIERS Normal Curve Equivalent Scores; Test Batteries; *Title I Evaluation and Reporting System

ABSTRACT

This study compares test battery results through carefully planned analysis of actual Title I evaluation data. The goal of the study is to derive practical recommendations regarding the results of commonly used test batteries which would be helpful at the local level. A criticism of the comparability assumption underlying the Title I evaluation and reporting system (TIERS) is not intended, rather, comparability within the current system with hopes of deriving useful recommendations is investigated. Six data sets from four Midwestern states from 1979 and 1980 were used. Data from reading and math programs, grades two through six, were included. Finding of this study include: (1) very few test batteries are actually given to 80% or more of the students; (2) normal curve equivalents from the Stanford Diagnostic Reading Test 76 and the Stanford Diagnostic Math Test 76 are often erroneously high because of incorrect conversion of Spring test scores; and (3) the 1978 edition of the Gates Mac Gintil Test appears to yield larger gains than the 1964-65 edition. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED208035

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The Consistency of Title I Evaluation Results for Commonly Used Test Batteries

Pat Thompson
American Institutes for Research

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. Thompson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 732

A paper presented at the American Educational Research Association annual meeting in Los Angeles, April 13-17, 1981.

ACKNOWLEDGMENTS

I would like to extend a special thanks to the people who made this study possible. First of all, a very special thanks to all the people in the State Title I Offices who generously let me use their data. Without their cooperation and interest this study would not have been possible. Secondly, thanks are extended to Judy Steenson for her assistance in keeping results of the study organized and for typing this paper. Thanks are also extended to Joe Buckley for sharing his knowledge of the literature in this area, and to Deanne Wehnes for her assistance with the many figures.

INTRODUCTION

The Title I evaluation and reporting system (TIERS) is based on the assumption that commonly used standardized tests are sufficiently comparable to allow aggregation of scores across tests. In fact, section 151 of P.L. 93-380 requires that objective criteria shall be specified and utilized in the evaluation of all programs which produce "data which are comparable on a statewide and nationwide basis." Yet, from the local level to the state level to the national level, there is little consistency in the choice of standardized tests for eveluation purposes. Instead a large variety of tests are used in Title I evaluation under the assumption that they yield comparable results.

An alternative to the present use of a variety of instruments in Title I evaluation would be to limit the choice to a single, system-wide test battery or to a few test batteries empirically shown to be comparable. Tallmadge and Horst (1978) discussed the comparability problem with 54 State Title I Coordinators and/or their designated evaluators to get their reaction to this alternate idea. Most of the States indicated that some degree of standardization would be desirable, but were hesitant regarding specific details. Apparently the States were resistant to the idea of change, not so much to the idea of standardization (pg. 2).

Tallmadge and Horst devote most of the paper to the content that, although comparability of results using different tests may suffer, it is more important not to limit test selection at the local level. Their argument is that the flexibility of the current system concerning test selection "actually <u>increases</u> the validity of local, State, and national Title I evaluations" (pg. 3). Assuming that local districts do in fact

select tests with content relevant to their programs, it is probably better to maintain flexibility in the system than to require standardization. Tallmadge and Horst are convinced that the current system yields fairer assessments of Title I accomplishments than would requiring a single test (pg. 12).

It is important to consider ramifications of the "current system" vs "standardization" issues regarding achievement testing in Title I evaluation. But it is also important to empirically study whether results are in practice comparable for students who took different test batteries and for students who took the same test battery. It may be that in practice results are comparable across test batteries. Or it may be that results are not even comparable for groups of similar students taking the same test battery. In the first case there is evidence to support the comparability assumption. In the second case, comparability of results across tests is no longer an issue.

Unfortunately there are few empirical studies on the comparability, equivalence, or difficulty level of different achievement test batteries, even during the last 30 year period. After exhaustive manual and computer searches on this topic Buckley (1981) states:

> The overall conclusion that can be reached is that comparative test difficulty level has received scant attention in the research literature. Articles that deal with difficulty level usually address the concept of the difficulty level of single test items rather than the achievement battery as a whole. (pg. 9)

Buckley came to this conclusion after thorough searches of Information Clearinghouse (ERIC), Psychological Abstracts, Social Science Citation Index, the Mental Measurements Yearbook (Buros, 1978), and various other bibliographies and newsletters on testing (pgs. 8-9).

The scarcity of literature on test difficulty and comparability is surprising given the magnitude of achievement testing in the United States. Based on 1976 data from the Government Accounting Office, 90% of the local educational agencies in the country administer norm referenced, standardized tests. In addition, two hundred million achievement test forms are used each year in the United States (Yeh, 1978).

The purpose of this study is to compare test battery results through carefully planned analyses of actual Title I evaluation data. The ultimate goal of the study is to derive practical recommendations regarding the results of commonly used test batteries which would be helpful at the local level. The purpose of the study is not to be critical of the comparability assumption underlying TIERS, but to investigate comparability within the current system in the hopes of deriving useful recommendations.

Of course, recommendations such as these which are generalizable can only result from a large study which has carefully been conceived. An effort was made in this study to include as much data as possible from many states and from the last two years, and also to carefully select data from these data sets for analysis.

This paper reports the results of descriptive analyses on the comparability of test results and is meant to be the starting point for more sophisticated statistical analyses to be reported later.

DESCRIPTION OF DATA USED IN THE STUDY

A total of six data sets were used in the study. The data sets contained Title I evaluation data submitted by local districts to their state Title I Office. The data were generously provided by four states in the Midwest and were from FY1979 and FY1980. It was important to employ data both across states (i.e. location) and across different years (i.e. time). If there are any consistency in test battery results, then they are most likely evident across time for the same location, or across locations for the same time. This should be true even given the large amount of data aggregated for each of the six data sets.

The size of the six data sets ranged from containing data for about 21,000 students to about 98,000 students. The average amount of data for the data sets combined was over 42,000 student records per set; the median was about 34,000 records. However, only a relatively small subset of the available data was used for this study. All of the data sets included data for almost all Title I students in the state during the year with the exception of the smallest data set which contained about 2/3 of all possible student data.

Most of the data sets consisted of individual student records with the exception of one data set which consisted of records with data aggregated across students. The data sets also differed in the ways raw test scores were converted into percentiles and then were converted into NCE scores[1]. For one data set, NCEs were submitted directly by the local

[1] NCEs (Normal Curve Equivalents) are used in the TIERS to measure student achievement and gain. The NCE scale is like a percentile scale in that it has values of "1" and "99" at the extremes and "50" in the middle. However the NCE scale is different in that it is an equal interval scale with equal distances (intervals) between individual points on the scale.

district. For two other data sets, raw scores were submitted by the district and a sophisticated computer program translated the raw scores into NCE scores via the test information provided and test norm table files. For the remaining four data sets, the districts submitted percentile scores which were then simply converted to NCE scores. Regardless of the record types and the means by which NCE scores were calculated, all data sets included information for the variables of descriptive interest in this study.

Data included in the second phase of analyses were carefully selected based on several important project and testing characteristics. One aim of the study was to analyze only the data meeting stringent requirements in terms of project and testing characteristics. Using only data meeting these requirements resulted in more controlled and homogeneous sets of data from which more precise generalizations could be made.

In terms of project characteristics, data from only pull-out reading or math programs in grades two through six were included. Reading and math data were always analyzed separately. In addition, only data for students who were in the Title I program for at least 50 instructional hours were included in the analyses. Table 1 shows that all four project characteristics were used to select data from all data sets except one where some information was not available.

Data were also selected based on the degree to compliance with several requirements regarding testing. First of all, only student data where a gain score could be calculated were included. This requirement implies that the student had to be in the Title I program at least six months. Secondly, only data from a fall-to-spring testing interval were selected.

TABLE 1: PROJECT AND STUDENT CHARACTERISTICS USED IN SELECTING DATA FROM SIX DATA SETS

	DATA SET [c]					
	A-79	A-80	B-79	B-80	C-80	D-79
PROJECT CHARACTERISTICS						
1. Reading and Math programs	X	X	X	X	X[a]	X
2. Grades 2 through 6	X	X	X	X	X	X
3. Only pull-out programs	X	X	X	X		X
4. At least 50 total hours of instructional time	X	X	X	X		X
TESTING CHARACTERISTICS						
5. Have a gain score (so took both tests)	X	X	X	X	X	X
6. Fall to Spring test interval	X	X	X	X	X	X
7. Same test pre to post	X	X	X	X	X[b]	X
8. Same subtest pre to post	X	X	X	X		X
9. Tested within 30 absolute days of norming dates	X	X	X	X		X
10. Not selected using pretest			X	X		X
11. Both tests given inlevel				X		

[a] Data for Language Arts students were included with Reading student data.

[b] Assumed same test was given but could not be empirically checked.

[c] The letters A to D represent the four states while 79 and 80 represent FY1979 and FY1980 respectively.

Research has repeatedly shown that test results vary depending on the testing interval used, so testing interval was held constant. Other testing requirements were that the same test and subtest was given for the pretest and posttest to help ensure comparability. Fifth, the two testings must have been within 30 total absolute days of the appropriate norm dates. One implementation rule in Title I evaluation is that test be given within a period $\pm$ two weeks from the empirical norm dates. Data were selected from all but one data set based on all five of these criteria. Finally, data were omitted from analysis if the pretest was used for selection and/or if out-of-level testing was used when it was possible to use these criteria as well. Table 1 displays which testing characteristics were used in selecting data from each of the data sets. Again, the purpose of selecting data meeting a well devined set of project and testing characteristics was to analyze carefully defined and more homogeneous data. Many studies using Title I evaluation data fail to control for variables not of interest in the study but which compound the findings beyond meaningfulness.

FREQUENCY OF TEST BATTERY USAGE

The first phase of analyses were run on reading and math data separately to determine which test batteries were given for Title I evaluation purposes. These descriptive analyses included all data for students in grades two through six who had gain scores. The analyses provided descriptive information regarding test usage and determined which tests would be looked at more closely in the second phase of analyses.

Table 2 shows the cummulative precent of students according to the number of test batteries taken. For example, only five or six test batteries explain over 80% of the reading tests taken for each individual data set. In math, from four to seven test batteries were taken by at least 80% of the math students depending upon the data set. These results indicate that although the local districts could use any of a large number of tests in the states providing data for this study, a relatively small number of tests are actually given to most students.

The test batteries given to at least 80% of the students from each data set are listed in Table 3. The table contains the percentage of students who took the test battery and the test rank within the data set. A composite across data sets is not given because the states providing data have unequal numbers of data sets. Still, some commonality in test battery usage across the states and the two years can be seen.

Excluding cases where it was uncertain which test edition was given, a total of 11 different test batteries were given to 80% or more reading students within the six data sets. The most frequently used tests for reading students were the 1964/5 and 1978 editions of the Gates-MacGinitie Reading Test (Gates). Other frequently used tests were the 1970 edition of the Metropolitan Achievement Test (MAT), the Stanford Achievement Test (SAT 73), and the Stanford Diagnostic Reading Test (SDRT 76).

The tests most frequently given to math students were the Stanford Diagnostic Math Test (SDMT 76), the SAT 73, and the 1970 and 1978 editions of the MAT. Another commonly used test in math was the California Achievement Test, 1977/8 edition (CAT 77/8). The SRA Achievement Tests (SRA) were the top ranking tests used for math in one state during two years, however

TABLE 2: Cummulative Percent of Students According to the Number of Test Batteries Taken[a]

READING

DATA SET	Number of Test Batteries						
	one	two	three	four	five	six	seven
A-79	19.8	38.5	55.9	71.6	79.2	85.2	
A-80	26.5	43.9	55.9	66.2	75.1	81.4	
B-79	32.5	53.8	66.3	76.7	83.7		
B-80	48.0	57.9	67.0	74.8	81.4		
C-80	24.3	46.6	61.2	72.6	79.9	85.4	
D-79	29.2	48.5	62.3	68.8	74.8	80.7	

MATH

A-79	35.6	58.1	72.4	82.3			
A-80	28.5	44.9	56.9	64.5	71.5	77.9	82.3
B-79	25.3	46.2	65.7	78.4	90.4		
B-80	27.2	44.8	61.0	76.3	90.2		
C-80	31.0	52.3	66.8	77.7	83.2		
D-79	29.0	47.9	59.8	67.9	75.6	81.6	

[a]These data are based only on students in grades two through six who had gain scores.

TABLE 3: TEST BATTERIES MOST COMMONLY USED IN GRADES TWO THROUGH SIX BY DATA SET[a]

READING TESTS	DATA SET											
	A-79		A-80		B-79		B-80		C-80		D-79	
	rank	%	rank	%	rank	%	rank	%	rank	%	rank	%
CAT[b]									1	24.3%		
CAT 77/8			5	8.9%								
Gates[b]									2	22.3%		
Gates 64/5	1	19.8%	3	12.0%	1	32.5%	6	6.1%			2	19.3%
Gates 78	3	17.4%	1	26.5%	2	21.3%	1	48.0%			4	6.5%
ITBS 71											1	29.2%
MAT[b]									6	5.5%		
MAT 70	4	15.7%	6	6.3%	3	12.5%	3	9.1%				
MAT/S 78					5	7.0%	2	9.9%				
SAT 73	6	6.0%			4	10.4%	4	7.8%	3	14.6%	3	13.8%
SDRT 76	5	7.6%	4	10.3%			5	6.6%	5	7.3%	5	6.0%
SRA[b]	2	18.7%	2	17.4%								
SRA ACH 71											7	3.8%
SRA ACH 78											6	5.9%
Woodcock									4	11.4%		

TABLE 3: Continued

MATH TESTS	DATA SET											
	A-79		A-80		B-79		B-80		C-80		D-79	
	rank	%	rank	%	rank	%	rank	%	rank	%	rank	%
CAT[b]									1	31.0%		
CAT 77/8			7	4.4%	5	12.0%	4	15.3%			4	8.1%
ITBS 71											1	29.0%
Keymath									4	10.9%		
MAT[b]									5	5.5%		
MAT 70	2	22.5%	3	12.0%	1	25.3%	2	17.6%				
MAT 78			4	7.6%								
MAT/I 78											5	7.7%
MAT/S 78					3	19.5%	1	27.2%				
SAT 73	4	9.9%	6	6.4%	2	20.9%	3	16.2%	2	21.3%	2	18.9%
SDMT 76	3	14.3%	2	16.4%	4	12.7%	5	13.9%	3	14.5%	3	11.9%
SRA[b]	1	35.6%	1	28.5%								
SRA ACH 71											6	6.0%
WRAT			5	7.0%								

[a]These data are based on students who had gain scores.

[b]It was not possible to determine which edition of these tests were used so this information is kept separate.

it was not possible to determine which editions of the SRA were given.

THE CONSISTENCY OF TEST BATTERY RESULTS ACROSS LOCATION AND TIME

The second phase of analyses involved selecting data from each data set which met the project and testing requirements previously described. Data were also selected only where students took one of the most commonly used test batteries relative to other tests given within the state that year. Analyses were run separately for reading and math data from each data set.

Average pretest, posttest, and gain NCE scores were calculated by grade for particular subtests of each commonly used test battery. For most reading tests, average NCE scores were based only on student data from the Total Reading subtest. In a few cases where Total Reading scores were not available for all grades of focus, either Total Reading or Comprehension subtest scores were used. Only Total Math subtest scores were selected for the analyses of the math data. Analyses were limited to Total Reading, Reading Comprehension, or Total Math subtests because scores from these subtests are generally the most reliable. Again, in comparative analyses such as these, it is important to compare data selected using a common "yardstick".

Results from many of the frequently used test batteries dropped out of this second phase of analyses because of the stringent requirements used for data selection. For example, 46.9% of SDRT 76 scores were lost from one data set when student selection based on the pretest was used for omitting data. Thus, results based on a sufficient sample size were not available for all of the frequently used tests from each data set.

Reading

Analyses of reading test results by grade produced results with a sufficient sample size per cell ($N \geq 100$) from two or more data sets for the following five test batteries: 1) Gates 64/5, 2) Gates 78, 3) SDRT 76, 4) MAT 70, and 5) SAT 73. The consistency of test results for each of these five test batteries across data set and student grade is presented after a discussion of the results ignoring student grade level.

Test Battery Results Ignoring Grade. The consistency of mean NCE gains across data sets for five commonly used test batteries in reading is presented at the top of Table 4. Data from grades two through six are colapsed in this presentation. Instead, the emphasis is on looking for overall consistency across the various data sets for each test battery listed. These means are based on sample sizes ranging from 167 to 4,103 student records per cell with an average cell sample size of 1,040 (median N was 885). Pretest NCEs for the same students with gains reported in Table 4 are presented in Table 5. The mean pretest NCE scores give more meaning to interpreting the gains in Table 4 because they offer a common starting point for comparing test results. Although statistical analysis of the beginning pretest scores are not reported for this descriptive phase of the study, it appears that there are no systematic biases in the mean pretest NCEs as a function of test battery.

The mean NCE reading gains for the five test batteries are also presented in Figure 1. Inspection of Table 4 and Figure 1 indicate that overall mean NCE gains for the Gates 64/5 and Gates 78 reading tests are realtively stable across five different data sets from three states. In addition, overall means from the Gates 78 are usually higher than means

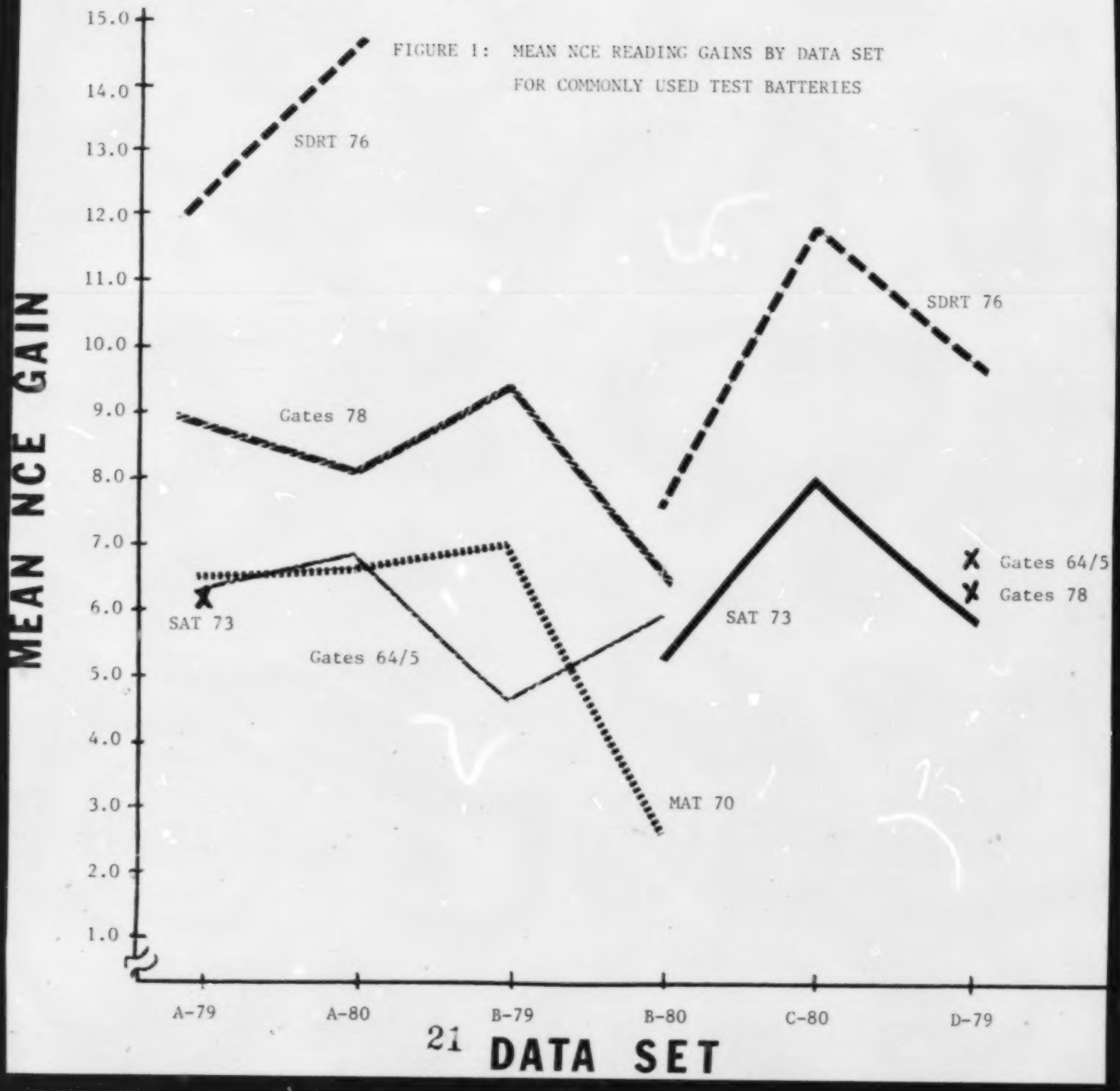
TABLE 4: MEAN NCE GAINS ACROSS SIX DATA SETS FOR THE MOST COMMONLY USED TEST BATTERIES

READING TESTS	DATA SET											
	A-79		A-80		B-79		B-80		C-80		D-79	
	$\bar{X}$	(SD)	$\bar{X}$	(SD)	$\bar{X}$	(SD)	$\bar{X}$	(SD)	$\bar{X}$	(SD)	$\bar{X}$	(SD)
Gates 64/5	6.25	(12.7)	6.79	(11.5)	4.57	(13.8)	5.81	(14.6)	---		6.81	( )
Gates 78	8.80	(12.0)	8.01	(11.8)	9.37	(12.3)	6.57	(10.7)	---		6.36	( )
MAT 70	6.43	(11.6)	6.58	(10.4)	6.96	(12.1)	2.56	(11.2)	---		---	
SAT 73	6.06	(9.8)	---		---		5.27	(10.0)	7.94	(12.2)	5.88	( )
SDRT 76	11.85	(14.0)	14.47	(16.5)	---		7.53	(12.4)	11.75	(13.6)	9.76	( )
MATH TESTS												
SDMT 76	15.78	(15.4)	14.21	(12.7)	5.20	(12.7)	7.40	(11.1)	11.68	(15.1)	10.52	( )
SAT 73	7.28	(10.7)	---		---		---		11.84	(13.8)	8.48	( )
MAT 70	9.41	(11.2)	7.47	(12.9)	---		---		---		---	
CAT 77/8	---		---		7.27	(13.3)	9.75	(12.2)	---		8.48	( )

TABLE 5: MEAN PRETEST NCEs ACROSS SIX DATA SETS FOR THE MOST COMMONLY USED TEST BATTERIES

READING TESTS	DATA SET											
	A-79		A-80		B-79		B-80		C-80		D-79	
	$\bar{X}$	(SD)	$\bar{X}$	(SD)	$\bar{X}$	(SD)	$\bar{X}$	(SD)	$\bar{X}$	(SD)	$\bar{X}$	(SD)
Gates 64/5	35.49	(11.4)	36.49	(11.0)	37.86	(15.4)	38.38	(15.0)	---		34.46	( )
Gates 78	33.62	(10.7)	35.06	(11.4)	36.63	(13.9)	36.63	(13.1)	---		30.43	( )
MAT 70	36.51	(10.8)	37.58	(9.0)	35.17	(13.2)	34.50	(18.0)	---		---	
SAT 73	39.54	(10.6)	---		---		37.93	(17.1)	31.40	(15.4)	36.52	( )
SDRT 76	37.02	(10.5)	35.92	(13.1)	---		39.98	(13.4)	36.46	(13.7)	37.59	( )
MATH TESTS												
SDMT 76	38.77	(13.5)	35.65	(13.2)	37.70	(11.7)	38.64	(11.5)	34.99	(13.4)	36.23	( )
SAT 73	34.27	(12.1)	---		---		38.25	(8.6)	28.25	(13.8)	37.21	( )
MAT 70	34.44	(10.4)	37.15	(9.5)	---		34.63	(18.8)	---		---	
CAT 77/8	---		---		44.13	(13.7)	44.90	(12.3)	---		32.27	( )

FIGURE 1: MEAN NCE READING GAINS BY DATA SET FOR COMMONLY USED TEST BATTERIES



from the earlier 1964/5 edition of the same test. This is true for two years of data from two states. Mean NCE reading gains are less stable across four data sets for the SAT 73 and MAT 70 tests.

Results from the SDRT 76 are quite unstable based on five data sets. However, the interesting finding regarding the SDRT 76 is that NCE reading gains are much higher for every data set when this test is compared with other tests. Although this appears to be an exciting finding with practical implications, a caution is in order. Spring norms for the SDRT 76 were published separately from the test manuals in 1978 "Supplementary Norms Booklets." As a result, many districts do not have spring norms for their SDRT 76 tests and use fall norms for both the fall and spring conversions. Thus, gains for many districts using the SDRT 76 are erroroneously high.[2]

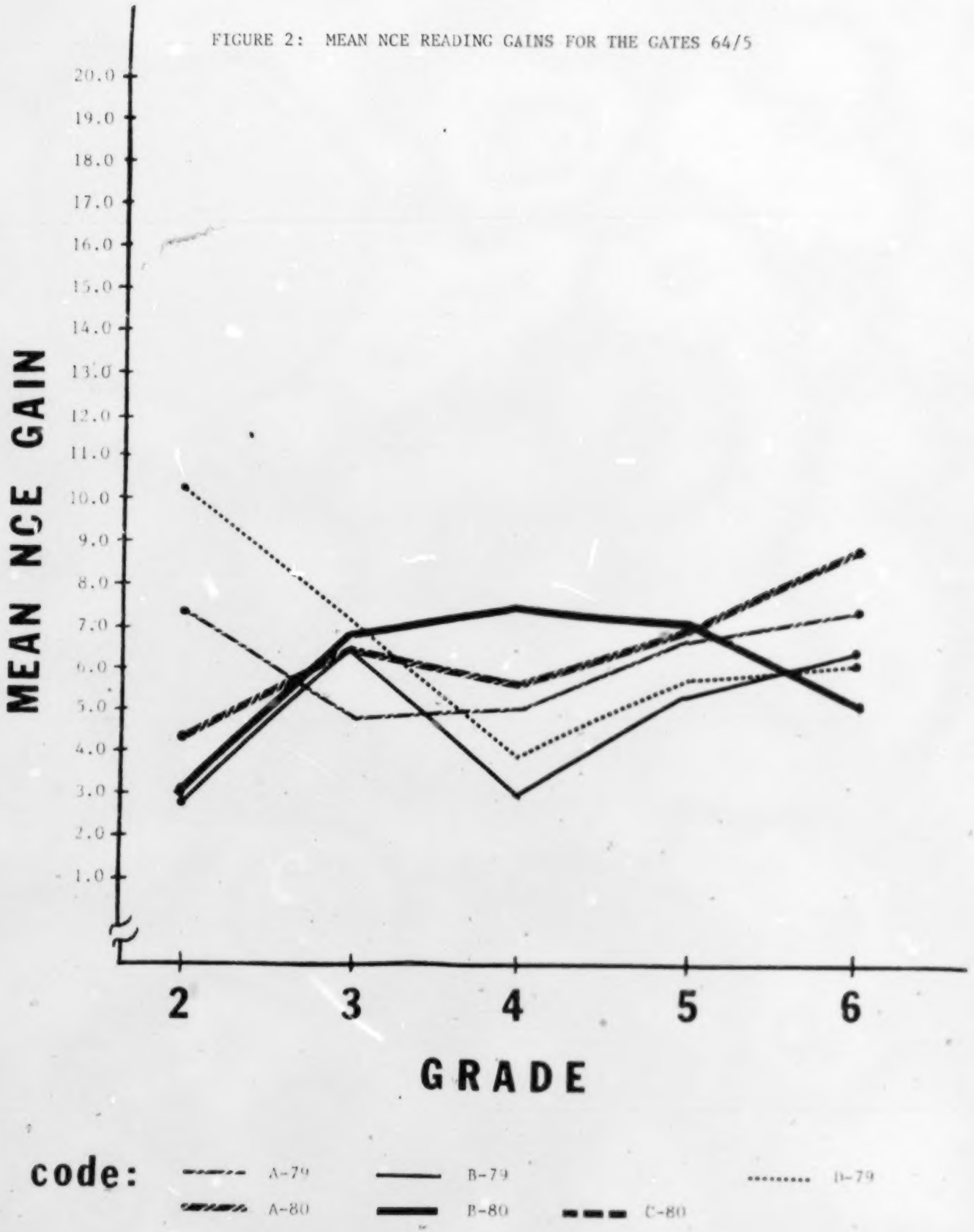
Gates MacGinitie, 1964/5. There was sufficient data[3] for each grade across five data sets (representing 3 states) for reading students who took the Gates 64/5 test. The mean NCE gains for the Gates 64/5 across grades 2 through 6 and for each data set are displayed in Figure 2. Mean

---

2 The author is indebted to Dr. Terry Armstrong for suggesting this possibility regarding an earlier study by Thompson & Novak (1980). When a subset of the data used for the study were checked manually for conversion errors, a large number of districts had in fact used fall norms for their spring conversions. In addition, there is evidence that SDRT 76 gains are not unusually high relative to other tests when fall and spring conversions are done by the computer.

3 The term "sufficient data" used in the context of the reading analyses refers to sample sizes per cell of typically at least 100 students. For example, all 25 cells for the Gates 78 analyses (5 grades X 5 data sets = 25) had sample sizes ranging from 141 to 462 with an average cell size of about 269 students. Out of the grand total of 85 cells reported for the second phase of reading analyses, only 4 cells had sample sizes slightly less than 100 but were not omitted from the study.

FIGURE 2: MEAN NCE READING GAINS FOR THE GATES 64/5



gains from the Gates 64/5 are fairly consistent across the five data sets for most grades except for grade 2. The average NCE gains from the Gates 64/5 are particularily consistent across data sets for grades 3 and 5. For grade 2, mean gains vary by as much as 7 NCEs for different data sets.

Gates MacGinitie, 1978. Sufficient data for each grade across five data sets (representing 3 states) were also available for the 1978 edition of the Gates. As can be seen in Figure 3, NCE gain scores for the 1978 edition are less stable across data sets than are gain scores from the earlier edition. However, NCE gains from the Gates 78 are quite consistent across sets for grades two and three. At grade four, a divergence begins where two FY1979 data sets have roughly equal larger gains than do three other data sets which are similar to one another. In other words, a two pronged type of diversion for gain scores begins at grade four, although results for the Gates 78 are fairly consistent for grades two and three.

Stanford Diagnostic Reading Test, 1976. A total of three data sets from three different states had sufficient SDRT 76 data at each grade. The mean NCE gains for the SDRT 76 are presented in Figure 4. The results across grades three through six are fairly stable and show a constant downward trend as grade level increased. The results at grade two, however, are remarkably unstable with a range of about 4.0 NCEs to 16.0 NCEs for the two extreme gain scores. Based on an earlier point made regarding the SDRT 76 scores used here (see Note 4) and the author's familiarity with the data sets involved, this finding is most likely due to incorrect score conversions.

Metropolitan Achievement Test, 1979 and Stanford Achievement Test, 1973. There was sufficient data from two data sets at each grade level for both

FIGURE 3: MEAN NCE READING GAINS FOR THE GATES 78

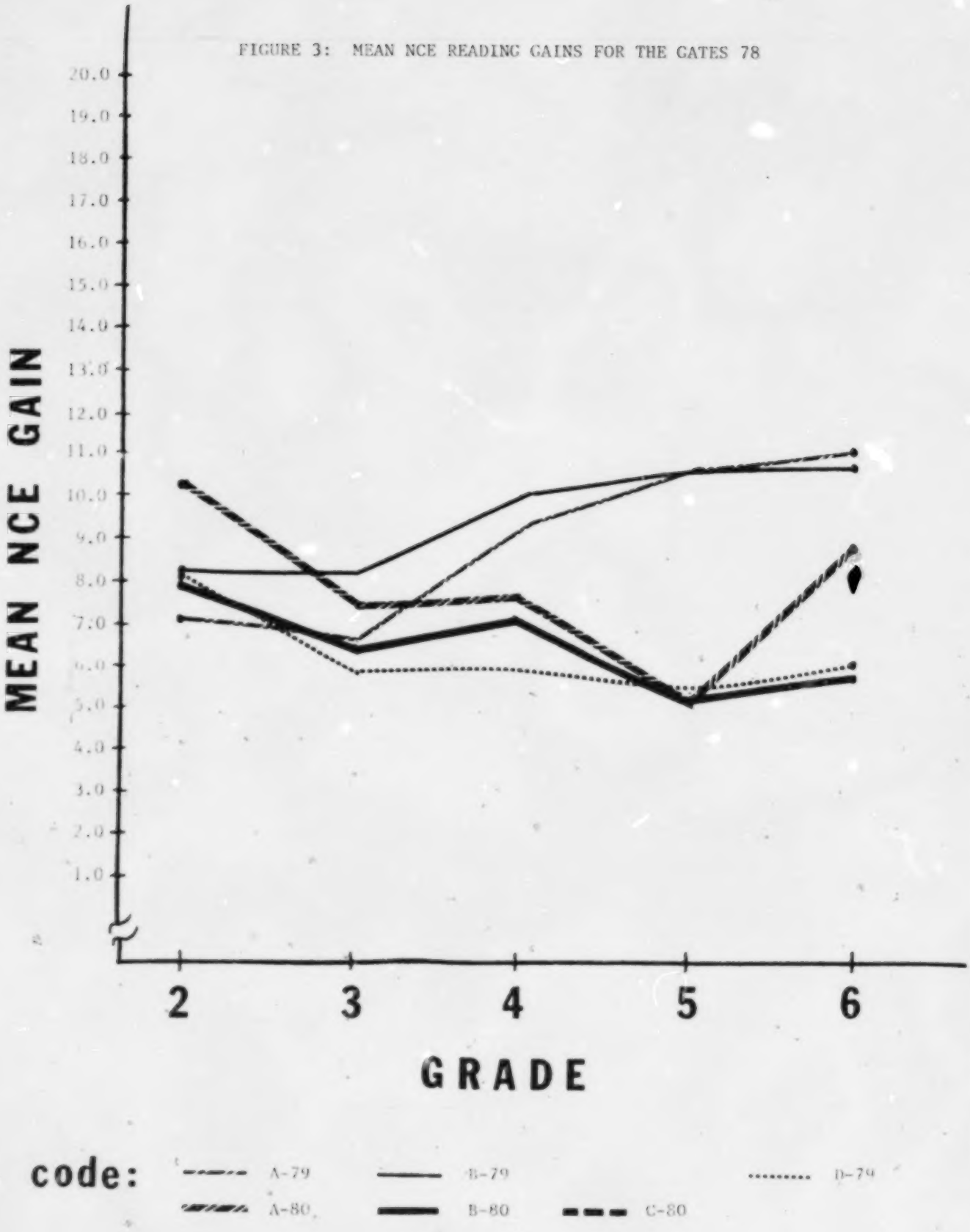
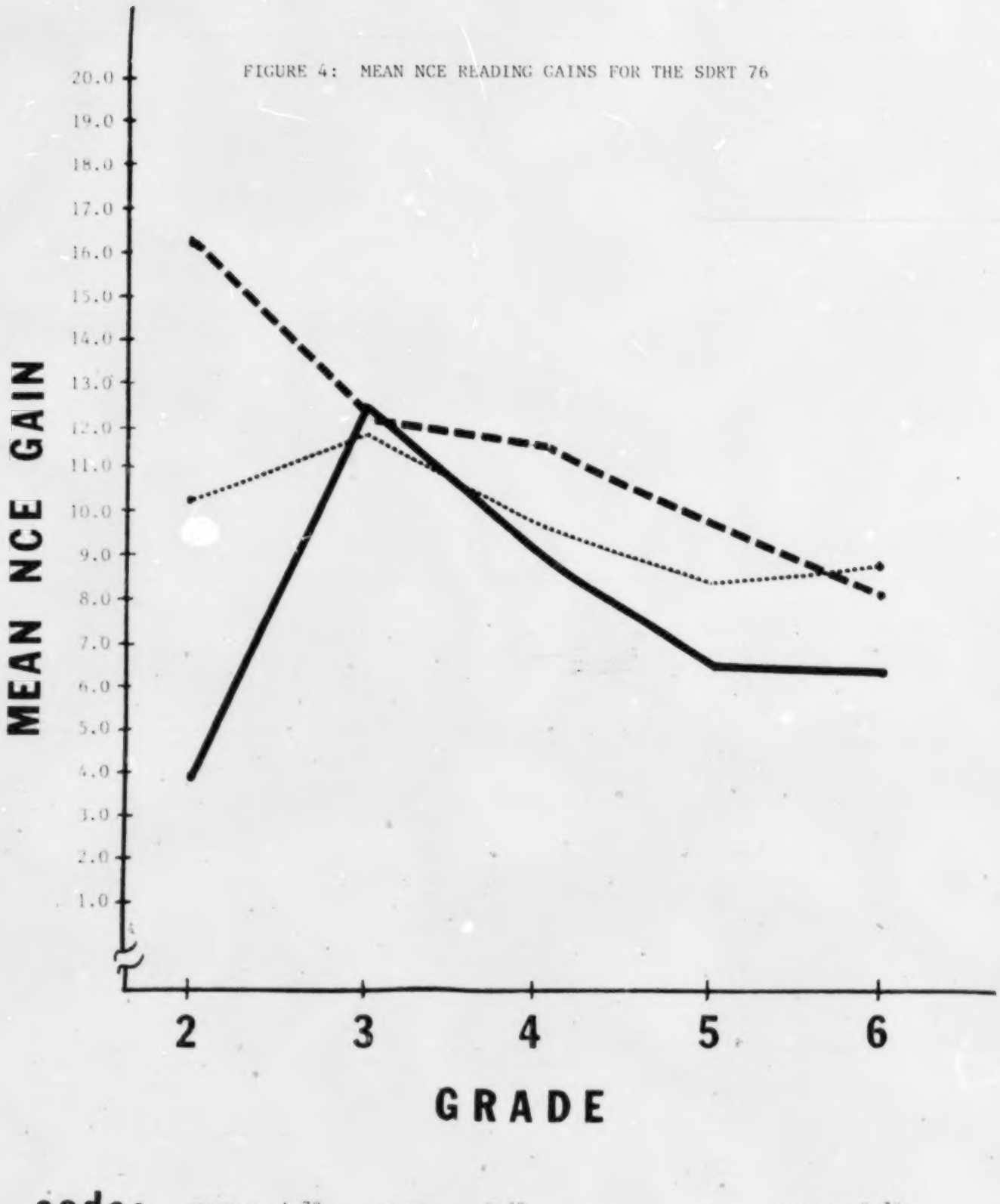


FIGURE 4: MEAN NCE READING GAINS FOR THE SDRT 76



of these tests. For the MAT 70 the two data sets were from one state and for the SAT 73 they were from two states. As can be seen in Figures 5 and 6, mean NCE gains across grades for the two data sets are not consitent with each other. For both of these tests, the largest descrepancies again are at grade two. The two data sets from the same state over two years yield somewhat similar results on the MAT 70, with the exception of the grade two means.

Math

Analyses of math tests results yielded small sample sizes for some grades by data cells, even for commonly used math tests. Even for the most frequently used math test, the SDMT 76, the sample sizes for the 20 cells ranged from only 38 to 267, although the average cell size was 116 students. Math test results from the SAT 73 and the CAT 77/8 will also be discussed here. However, average sample sizes for data sets were only 44 and 289 student records for the two data sets using the SAT 73 the most, and were only 65 and 168 records for the two data sets using the CAT 77/8 the most. The unequal cell sample sizes across different data sets and the small sample sizes, should be considered in interpreting the results of the math analyses.

Test Battery Results Ignoring Grade. The consistency of mean NCE gains across data sets for four commonly used math tests are presented at the bottom of Table 4 and in Figure 7. Data from grades two through six are colapsed in this presentation. Instead, the emphasis is on looking for consistency across the various data sets for frequently used test batteries. These means are based on sample sizes ranging from 165 to

FIGURE 5: MEAN NCE READING GAINS FOR THE MAT 70

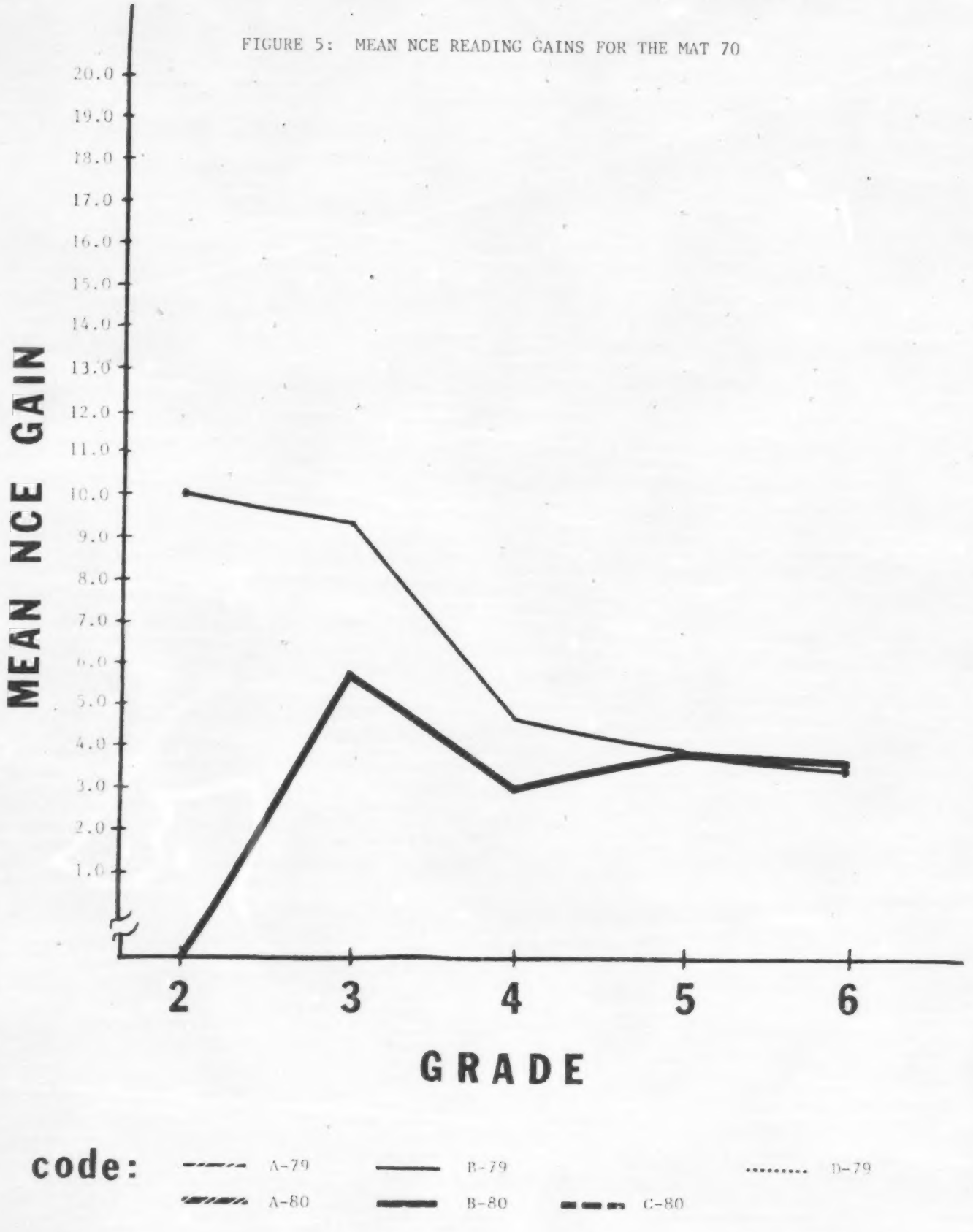


FIGURE 6: MEAN NCE READING GAINS FOR THE SAT 73

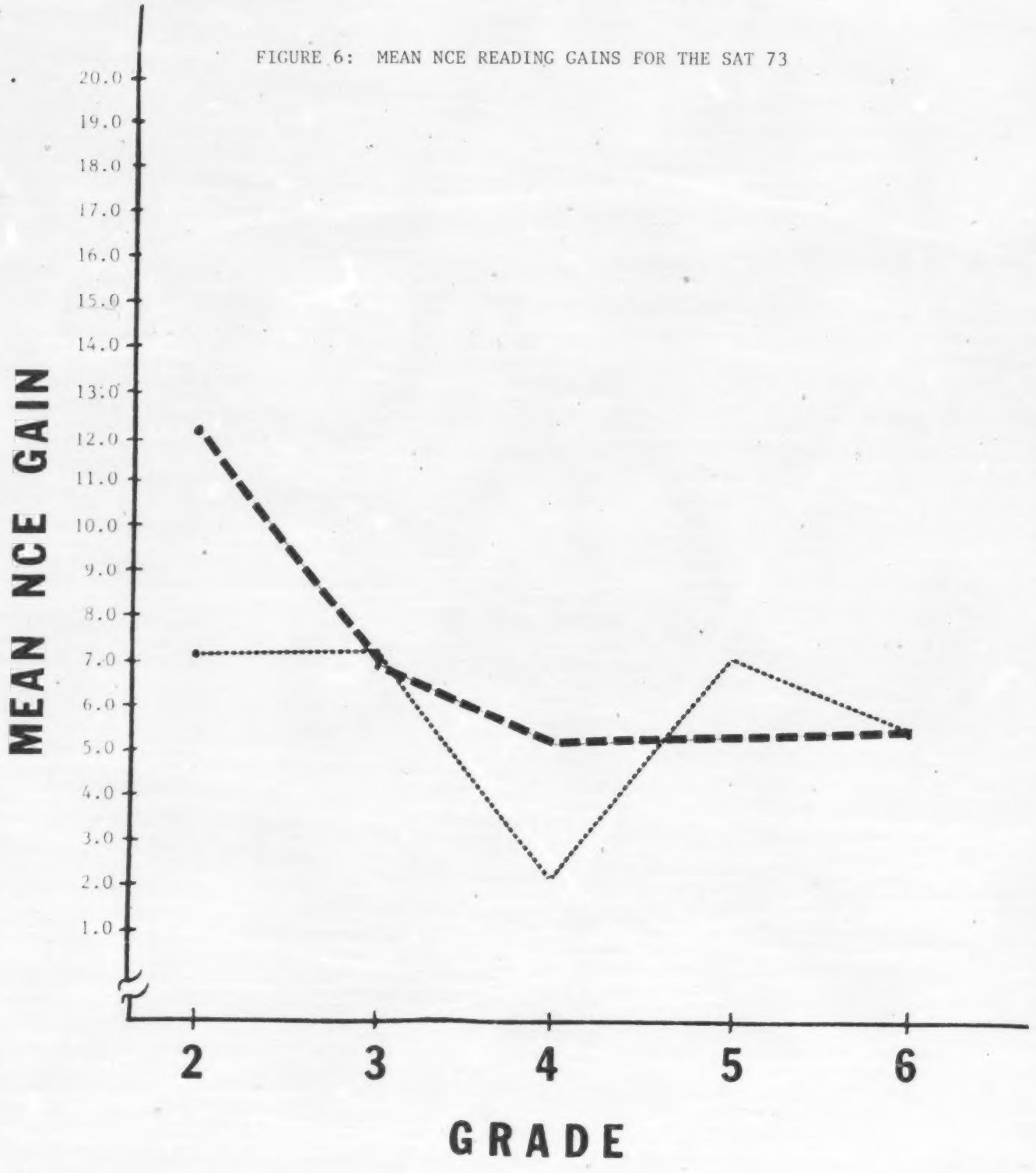
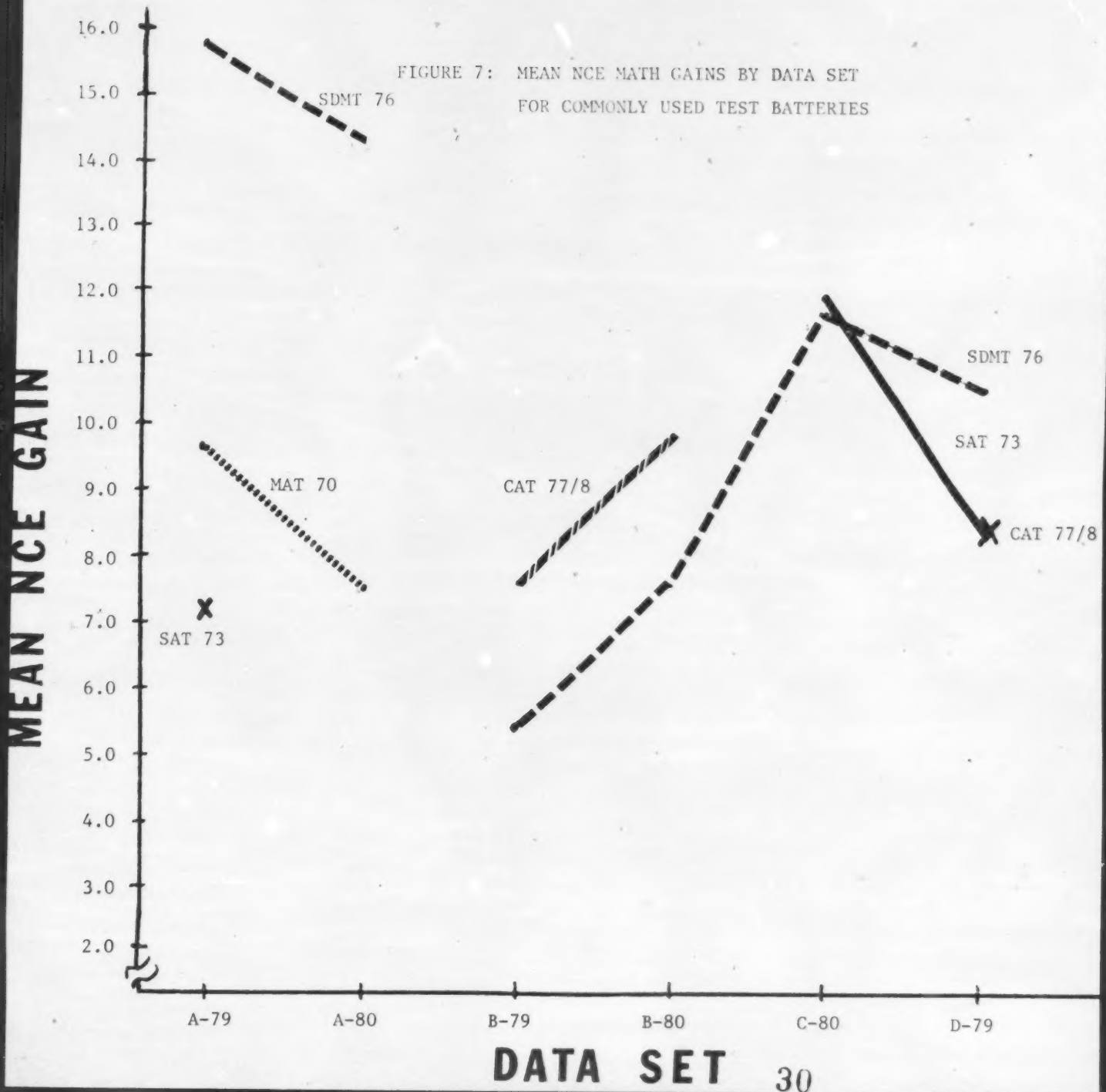


FIGURE 7: MEAN NCE MATH GAINS BY DATA SET FOR COMMONLY USED TEST BATTERIES



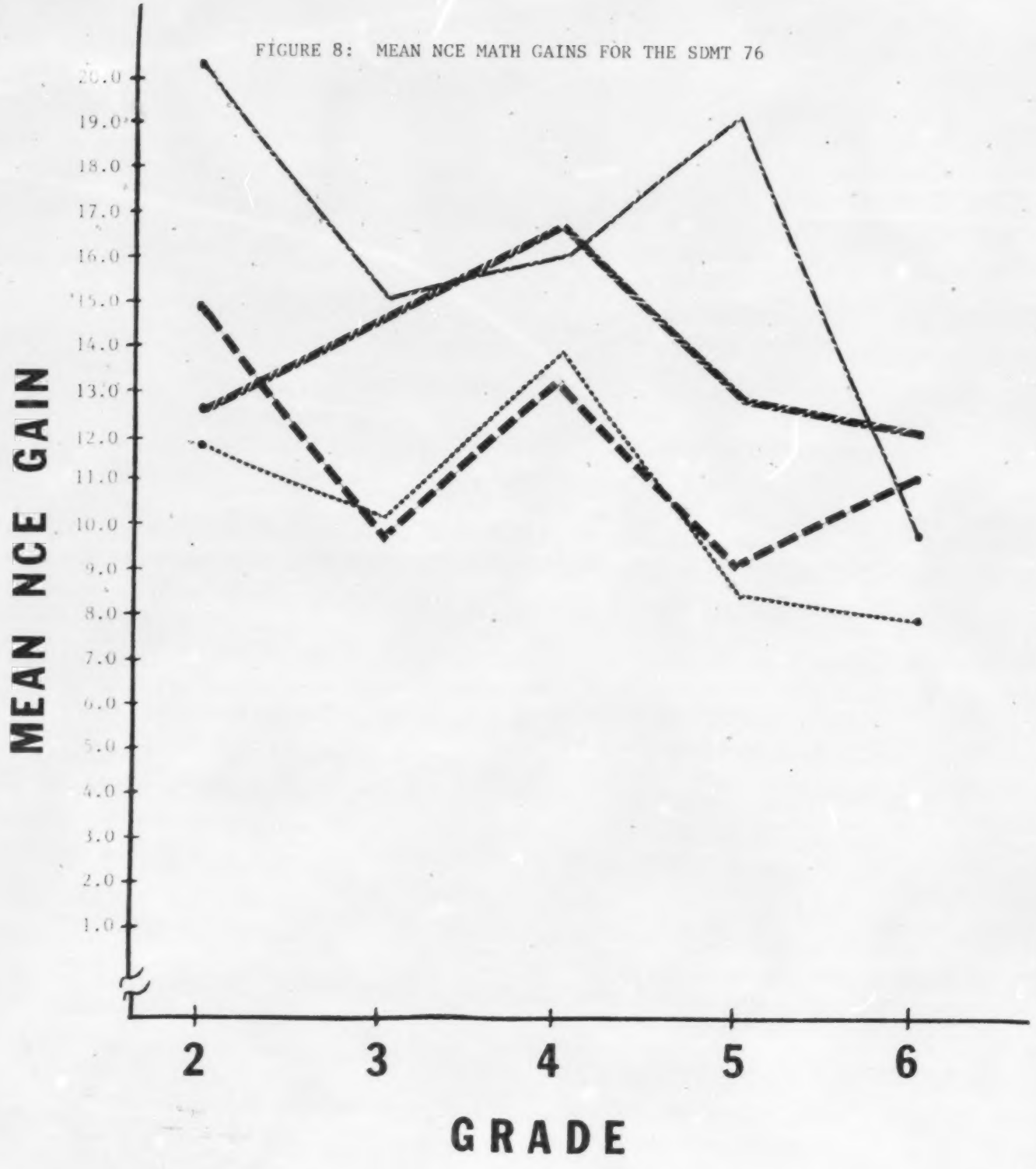
1946 student records per cell with an average cell sample size of 556 (median N was 320). Pretest NCE means for the same students with math gains are included in Table 5.

Inspection of Table 4 and Figure 7 indicate that overall mean NCE gains for the SDMT 76 are generally larger than those for the other three test batteries. It appears that problem previously discussed in reference to the SDRT 76 also apply to the SDMT 76. In fact, results from the SDMT 76 provide further evidence in support of the comments made in Note 2 at the bottom of page 4 .

Overall mean gains for the CAT 77/8 differed by about 1.3 NCEs for the two data sets from two states and two years. For the SAT 73, the average difference in NCEs for three data sets from three states was 3.0 NCEs. For the two data sets from the same state, the NCE difference on the MAT 70 was 1.9 NCEs. In general, overall mean NCE gains in math for the CAT 77/8, SAT 73, and MAT 70 were not widely different from each other even across four states and two years. In addition, overall means for the SDMT 76 would likely be similar for all data sets if the conversion of spring scores were done correctly.

Standord Diagnostic Math Test, 1976. Mean NCE gains by grade are plotted for four data sets (representing three states) who used the SDMT 76 in Figure 8. Although the gains are not really consistent across most data sets, one similar pattern does emerge. Average gains appear to be larger at grades two and four relative to smaller gains at grades three and five. This pattern was generally evident for three of the data sets from three states. Why this pattern emerged possibly could be explained by considering the relationship between grades and the levels of the SDMT 76. Unlike

FIGURE 8: MEAN NCE MATH GAINS FOR THE SDMT 76



many other tests, students in most grades may or may not be given different levels of the test from pre- to post testing. Either changing levels or using the same level is appropriate depending upon the discretion of the test administrator. The relatively larger gains for grades two and four may be primarily for students taking the same levels from pre to post testings where students from grades three and five may have changed to the higher level for the posttest. Thus, their gains would be relatively lower. In fact, for grade two it is not possible to change levels from pre to post testing.

Stanford Achievement Test, 1973. Mean gain results from two data sets from two states for the SAT 73 given to math students were not similar, as can be seen in Figure 9. For the data set with a very large sample size per cell, SAT 73 decreased slightly in a consistent manner as the grade level increased. However, for the data set with very small cell sample sizes, a "U" shaped function of mean scores across grades became evident.

California Achievement Test, 1977/8. As can be seen in Figure 10, a similar pattern of mean NCE gain results for two data sets from two states was found for the CAT 77/8 across all grades except one. Gains decreased sharply for grades two through four and increased for grade six. NCE mean gains in math differed quite a bit due to grade level for this test.

CONCLUSIONS

It appears that there are not large inconsistencies in average gain results due to using different test batteries, assuming that the use of incorrect norms tables is not systematic. At least the inconsistencies in test results found in this study do not appear limited to particular tests.

FIGURE 9: MEAN NCE MATH GAINS FOR THE SAT 73

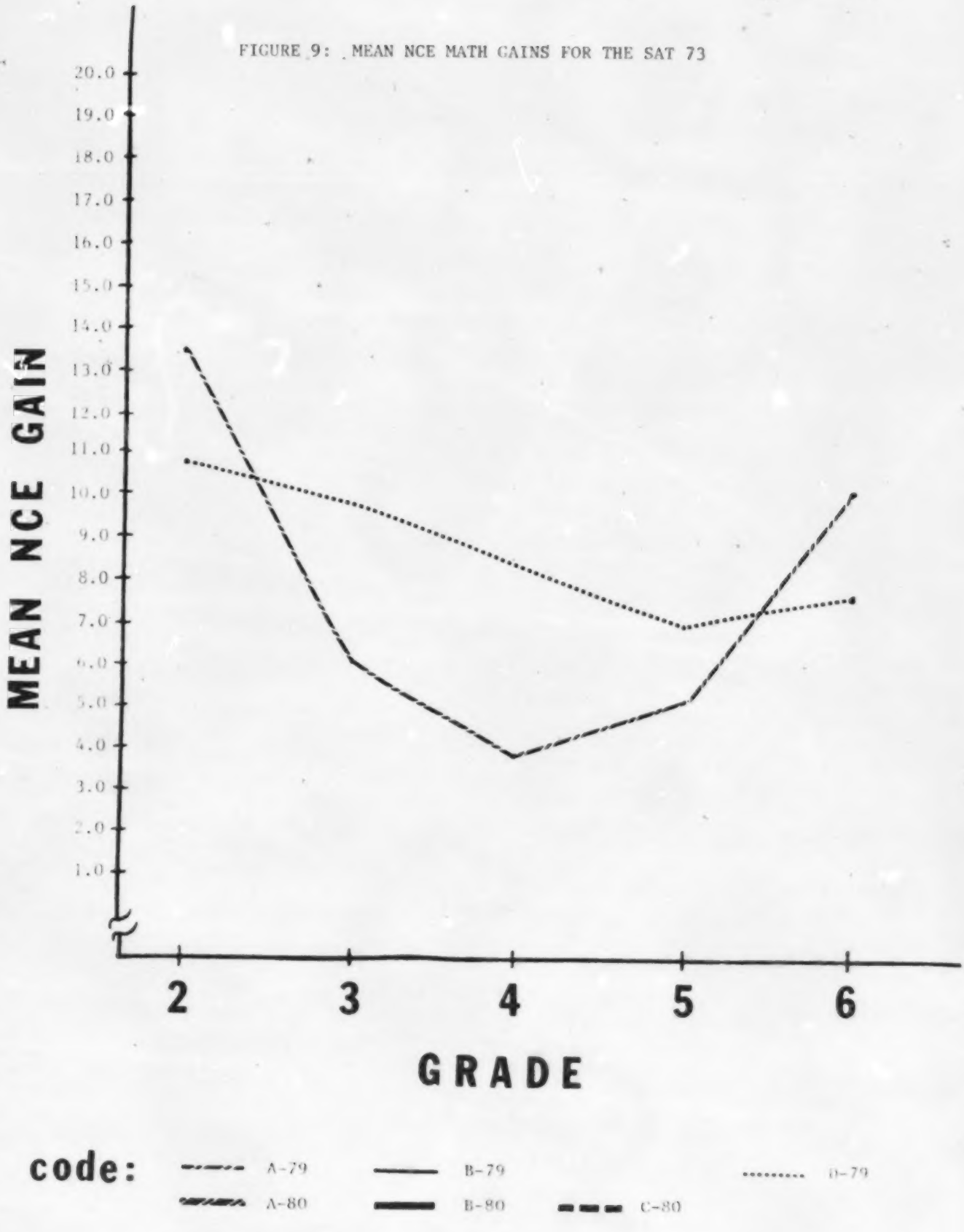
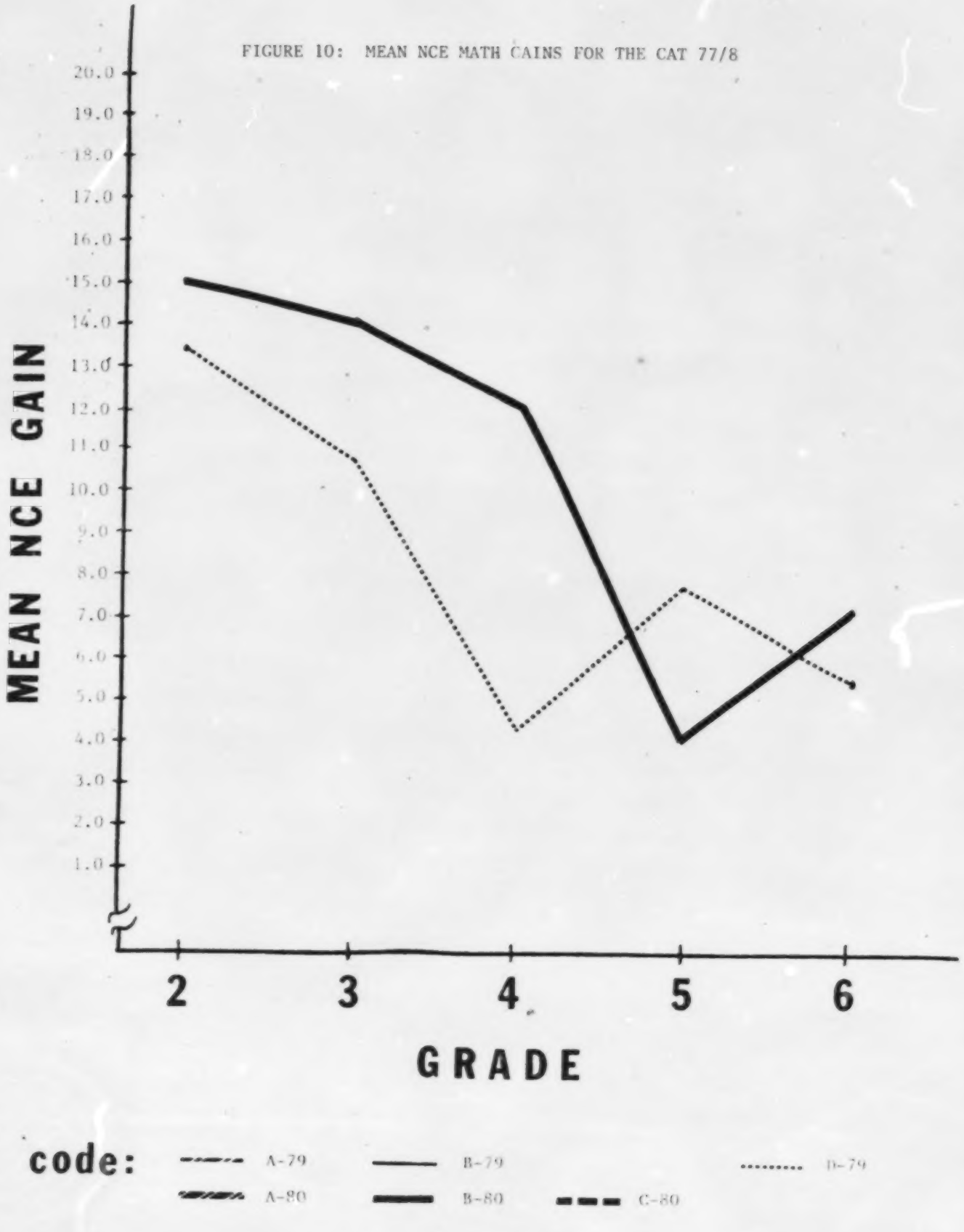


FIGURE 10: MEAN NCE MATH GAINS FOR THE CAT 77/8



At the same time, there does not appear to be overwhelming consistency in the results of different test batteries across location and time either. Without going beyond the descriptive analyses discussed in this paper to more controlled statistical analyses, it would be difficult at this point to make practical recommendations regarding the consistency of test battery results.

Some of the more interesting findings from this descriptive study are the following:

1) Very few test batteries are actually given to 80% or more of the students although numerous tests could be used,
2) NCE gains from the SDRT 76 and SDMT 76 are often erroneously high because of incorrect conversion of Spring test scores, and
3) The 1978 edition of the Gates appears to yield larger gains than the 1964/5 edition of the Gates.

The next step is to complete more sophisticated statistical analyses of test results for the most commonly used test batteries identified in this study. These analyses likely will be repreated measures analyses which focus on posttest differences controlling for entering pretest differences. The trends across both test batteries and grade levels will be investigated simultaneously. Even more carafully defined subsets of data may be selected for these analyses. The present study offers valuable descriptive information upon which the second phase of analyses now can be built.

REFERENCES

Buckley, Joseph James. A comparative analysis of the difficulty levels of three standardized achievement tests commonly used in ESEA Title I program evaluation. Unpublished master's thesis, California State University at Fresno, 1981.

Burós, Oscar K. The eight mental measurements yearbook (8th edition). Highland Park, New Jersey: Gryphon Press, 1978.

Tallmadge, G. Kasten and Horst, Donald P. The use of different achievement tests in the ESEA Title I evaluation system. Paper presented at the American Educational Research Association in Toronto, Canada, March, 1978.

Thompson, Pat and Novak, Carl D. The impact of testing decisions on Title I evaluation results. Paper presented at the Evaluation Research Society meeting in Washington, D.C., November, 1980.

Yeh, Jennie P. Test use in schools. Technical paper issued by the University of California, Los Angeles, June, 1978.